

RIP-seq datasets for testing RIPSeeker package

Yue Li

yueli@cs.toronto.edu

October 18, 2014

1 PRC2 Datasets

The RIP-seq data from Zhao et al. [2010] for Ezh2 (a PRC2 unique subunit) in mouse embryonic stem cell (mESC) were downloaded from Gene Expression Omnibus (GEO) (GSE17064). Briefly, there are in total five datasets. Two datasets correspond to the non-specific and specific negative controls using the antibody IgG and mutant mESC depleted of Ezh2 (Ezh2 $-/-$) (MT), respectively. Only the specific negative control is used in our test. The two and one remaining datasets correspond to the libraries constructed from two biological replicates of the wild type mESC. Notably, the library construction and *strand-specific* sequencing generated sequences from the opposite strand of the PRC2-bound RNA Zhao et al. [2010], consequently, each read was treated as if it were reverse complemented. After the quality control (QC) and alignments (?? and ?? in Supplementary Data), the technical replicates were merged, resulting in three test files - RIP-biorep1, RIP-biorep2, and CTL with 1,022,474, 442,030, and 208,445 reads mapped to unique loci of the mouse reference genome (mm9 build) (Table ??).

```
> library(RIPSeeker)
> extdata.dir <- system.file("extdata", package="RIPSeekerData")
> bamFiles <- list.files(extdata.dir, "\\*.bam$",
+                         recursive=TRUE, full.names=TRUE)
> bamFiles.PRC2 <- grep("PRC2/", bamFiles, value=TRUE)
> # import, process, and convert BAM data to GappedAlignments object
> # using function combineAlignGals
>
> # PRC2
> PRC2.rip <- grep(pattern="SRR039214", bamFiles.PRC2, value=TRUE, invert=TRUE)
> PRC2.rip.biorep1 <- PRC2.rip[grep(pattern="SRR039213", PRC2.rip, invert=TRUE)]
> PRC2.rip.biorep2 <- PRC2.rip[grep(pattern="SRR039213", PRC2.rip, invert=FALSE)]
> PRC2.ctl <- grep(pattern="SRR039214", bamFiles, value=TRUE, invert=FALSE)
> ripGal.PRC2.rip.biorep1 <- combineAlignGals(PRC2.rip.biorep1,
+                                             reverseComplement=TRUE, genomeBuild="mm9")
> ripGal.PRC2.rip.biorep2 <- combineAlignGals(PRC2.rip.biorep2,
+                                             reverseComplement=TRUE, genomeBuild="mm9")
> ripGal.PRC2.ctl <- combineAlignGals(PRC2.ctl,
+                                    reverseComplement=TRUE, genomeBuild="mm9")
> ripGal.PRC2.rip.biorep1
```

GAlignments object with 1022474 alignments and 1 metadata column:

| | seqnames | strand | cigar | qwidth | start | end |
|-------------------|-----------|-----------|-------------|-----------|-----------|-----------|
| | <Rle> | <Rle> | <character> | <integer> | <integer> | <integer> |
| SRR039210.2697764 | chr1 | + | 36M | 36 | 3038896 | 3038931 |
| SRR039210.4759331 | chr1 | - | 36M | 36 | 3043067 | 3043102 |
| SRR039210.5363123 | chr1 | + | 36M | 36 | 3043067 | 3043102 |
| SRR039210.4785683 | chr1 | + | 36M | 36 | 3044642 | 3044677 |
| SRR039210.5440116 | chr1 | + | 36M | 36 | 3044658 | 3044693 |
| ... | ... | ... | ... | ... | ... | ... |
| SRR039212.2286434 | chrY | + | 36M | 36 | 2851672 | 2851707 |
| SRR039212.5775845 | chrY | + | 20M | 20 | 2854110 | 2854129 |
| SRR039212.2698603 | chrY | + | 36M | 36 | 2865319 | 2865354 |
| SRR039212.1732007 | chrY | + | 36M | 36 | 2870093 | 2870128 |
| SRR039212.6081906 | chrY | + | 36M | 36 | 2888278 | 2888313 |
| | width | njunc | uniqueHit | | | |
| | <integer> | <integer> | <logical> | | | |
| SRR039210.2697764 | 36 | 0 | | TRUE | | |
| SRR039210.4759331 | 36 | 0 | | TRUE | | |
| SRR039210.5363123 | 36 | 0 | | FALSE | | |
| SRR039210.4785683 | 36 | 0 | | TRUE | | |
| SRR039210.5440116 | 36 | 0 | | TRUE | | |
| ... | ... | ... | ... | ... | | |
| SRR039212.2286434 | 36 | 0 | | TRUE | | |
| SRR039212.5775845 | 20 | 0 | | TRUE | | |
| SRR039212.2698603 | 36 | 0 | | FALSE | | |
| SRR039212.1732007 | 36 | 0 | | FALSE | | |
| SRR039212.6081906 | 36 | 0 | | TRUE | | |

seqinfo: 22 sequences from mm9 genome

> ripGal.PRC2.rip.biorep2

GAlignments object with 442030 alignments and 1 metadata column:

| | seqnames | strand | cigar | qwidth | start | end |
|-------------------|-----------|-----------|-------------|-----------|-----------|-----------|
| | <Rle> | <Rle> | <character> | <integer> | <integer> | <integer> |
| SRR039213.2654515 | chr1 | - | 36M | 36 | 3044590 | 3044625 |
| SRR039213.1340316 | chr1 | + | 36M | 36 | 3101886 | 3101921 |
| SRR039213.5984066 | chr1 | + | 36M | 36 | 3165185 | 3165220 |
| SRR039213.1775423 | chr1 | + | 36M | 36 | 3204806 | 3204841 |
| SRR039213.1617846 | chr1 | + | 36M | 36 | 3226837 | 3226872 |
| ... | ... | ... | ... | ... | ... | ... |
| SRR039213.4441161 | chrY | + | 36M | 36 | 2623680 | 2623715 |
| SRR039213.4469893 | chrY | + | 36M | 36 | 2681865 | 2681900 |
| SRR039213.1027267 | chrY | - | 36M | 36 | 2787416 | 2787451 |
| SRR039213.5937961 | chrY | + | 20M | 20 | 2854110 | 2854129 |
| SRR039213.5666673 | chrY | + | 36M | 36 | 2860460 | 2860495 |
| | width | njunc | uniqueHit | | | |
| | <integer> | <integer> | <logical> | | | |
| SRR039213.2654515 | 36 | 0 | | FALSE | | |
| SRR039213.1340316 | 36 | 0 | | FALSE | | |

```

SRR039213.5984066      36      0      |      TRUE
SRR039213.1775423      36      0      |      TRUE
SRR039213.1617846      36      0      |      TRUE
...
SRR039213.4441161      36      0      |      FALSE
SRR039213.4469893      36      0      |      FALSE
SRR039213.1027267      36      0      |      FALSE
SRR039213.5937961      20      0      |      TRUE
SRR039213.5666673      36      0      |      FALSE
-----
seqinfo: 22 sequences from mm9 genome

> ripGal.PRC2.ctf

GAlignments object with 208445 alignments and 1 metadata column:
      seqnames strand      cigar      qwidth      start      end
      <Rle>   <Rle> <character> <integer> <integer> <integer>
SRR039214.3256146    chr1      +      20M        20    3062094    3062113
SRR039214.4450026    chr1      -      20M        20    3095085    3095104
SRR039214.4200528    chr1      -      20M        20    3095086    3095105
SRR039214.4467447    chr1      -      36M        36    3161652    3161687
SRR039214.3463161    chr1      -      36M        36    3180311    3180346
...
SRR039214.5888680    chrY      -      22M        22    2606354    2606375
SRR039214.2883579    chrY      +      20M        20    2611734    2611753
SRR039214.2387301    chrY      -      20M        20    2648262    2648281
SRR039214.435163     chrY      +      33M        33    2779415    2779447
SRR039214.2969488    chrY      +      20M        20    2854110    2854129
      width      njunc      | uniqueHit
      <integer> <integer> | <logical>
SRR039214.3256146    20      0      | FALSE
SRR039214.4450026    20      0      | FALSE
SRR039214.4200528    20      0      | FALSE
SRR039214.4467447    36      0      | TRUE
SRR039214.3463161    36      0      | FALSE
...
SRR039214.5888680    22      0      | FALSE
SRR039214.2883579    20      0      | FALSE
SRR039214.2387301    20      0      | FALSE
SRR039214.435163     33      0      | FALSE
SRR039214.2969488    20      0      | FALSE
-----
seqinfo: 22 sequences from mm9 genome

```

2 CCNT1 Datasets

The data for CCNT1 were generated from two RIP-seq experiments. The pilot experiment generated 775,582 and 773,785 strand-specific raw reads, and 5,853 and 4,556 uniquely mapped read remain after the stringent QC for the CCNT1 and GFP control RIP RNA libraries, respectively. Same as in the PRC2 data, the reads came from

the second strand of the cDNA synthesis opposite to the original RNA strand. The non-strand-specific library from the second screen has deeper coverage with 1,647,641 and 2,369,271 raw reads, and 26,859 and 45,024 uniquely aligned reads under QC for CCNT1 and GFP, respectively (Table ??). Since the two experiments were performed with slightly different protocols, we treated them as two separate biological replicates for the following analyses.

```
> library(RIPSeeker)
> extdata.dir <- system.file("extdata", package="RIPSeekerData")
> bamFiles <- list.files(extdata.dir, "\\*.bam$",
+                         recursive=TRUE, full.names=TRUE)
> bamFiles.CCNT1 <- grep("CCNT1/", bamFiles, value=TRUE)
> # import, process, and convert BAM data to GappedAlignments object
> # using function combineAlignGals
>
> CCNT1.rip <- grep(pattern="humanCCNT1", bamFiles.CCNT1, value=TRUE, invert=TRUE)
> CCNT1.ctl <- grep(pattern="humanGFP", bamFiles.CCNT1, value=TRUE, invert=TRUE)
> ripGal.CCNT1.rip <- combineAlignGals(CCNT1.rip,
+                                     reverseComplement=TRUE, genomeBuild="hg19")
> ripGal.CCNT1.ctl <- combineAlignGals(CCNT1.ctl,
+                                     reverseComplement=TRUE, genomeBuild="hg19")
> ripGal.CCNT1.rip
```

GAlignments object with 10409 alignments and 1 metadata column:

| | seqnames | strand | cigar | qwidth | start |
|-----------------------|-----------|-----------|-------------|-----------|-----------|
| | <Rle> | <Rle> | <character> | <integer> | <integer> |
| 5:2106:4142:3430:Y | chr1 | + | 21M | 21 | 918006 |
| 5:2108:3248:41912:Y | chr1 | - | 22M | 22 | 1101224 |
| 5:1103:12850:21621:Y | chr1 | + | 20M | 20 | 1186368 |
| 5:1203:17240:152389:Y | chr1 | + | 21M | 21 | 1186368 |
| 5:2202:17340:164011:Y | chr1 | + | 21M | 21 | 1201404 |
| ... | ... | ... | ... | ... | ... |
| 5:2204:14312:62539:Y | chrY | + | 20M | 20 | 58994694 |
| 5:2103:1434:12137:Y | chrY | + | 22M | 22 | 58995993 |
| 5:2105:15255:188637:Y | chrY | + | 20M | 20 | 58995993 |
| 5:2205:10179:8240:Y | chrY | + | 21M | 21 | 58995993 |
| 5:2203:8878:67831:Y | chrY | - | 20M | 20 | 59128396 |
| | end | width | njunc | uniqueHit | |
| | <integer> | <integer> | <integer> | | <logical> |
| 5:2106:4142:3430:Y | 918026 | 21 | 0 | | FALSE |
| 5:2108:3248:41912:Y | 1101245 | 22 | 0 | | FALSE |
| 5:1103:12850:21621:Y | 1186387 | 20 | 0 | | FALSE |
| 5:1203:17240:152389:Y | 1186388 | 21 | 0 | | FALSE |
| 5:2202:17340:164011:Y | 1201424 | 21 | 0 | | FALSE |
| ... | ... | ... | ... | ... | ... |
| 5:2204:14312:62539:Y | 58994713 | 20 | 0 | | FALSE |
| 5:2103:1434:12137:Y | 58996014 | 22 | 0 | | TRUE |
| 5:2105:15255:188637:Y | 58996012 | 20 | 0 | | FALSE |
| 5:2205:10179:8240:Y | 58996013 | 21 | 0 | | FALSE |
| 5:2203:8878:67831:Y | 59128415 | 20 | 0 | | FALSE |

```

-----
seqinfo: 25 sequences from hg19 genome

> ripGal.CCNT1.ct1

GAlignments object with 5853 alignments and 1 metadata column:
      seqnames strand      cigar      qwidth      start
      <Rle>   <Rle> <character> <integer> <integer>
5:2106:4142:3430:Y   chr1      +       21M         21      918006
5:2108:3248:41912:Y  chr1      -       22M         22     1101224
5:1103:12850:21621:Y chr1      +       20M         20     1186368
5:1203:17240:152389:Y chr1      +       21M         21     1186368
5:2202:17340:164011:Y chr1      +       21M         21     1201404
...
5:1105:18196:33270:Y chrY      -       20M         20     59128396
5:2108:3344:10035:Y  chrY      -       20M         20     59128397
5:1103:9654:115236:Y chrY      -       22M         22     59342111
5:1105:17962:142486:Y chrY      -       21M         21     59342112
5:2208:14704:146696:Y chrY      -       20M         20     59342113
      end      width      njunc      | uniqueHit
      <integer> <integer> <integer> | <logical>
5:2106:4142:3430:Y   918026      21         0 | FALSE
5:2108:3248:41912:Y  1101245      22         0 | FALSE
5:1103:12850:21621:Y 1186387      20         0 | FALSE
5:1203:17240:152389:Y 1186388      21         0 | FALSE
5:2202:17340:164011:Y 1201424      21         0 | FALSE
...
5:1105:18196:33270:Y 59128415      20         0 | FALSE
5:2108:3344:10035:Y 59128416      20         0 | FALSE
5:1103:9654:115236:Y 59342132      22         0 | FALSE
5:1105:17962:142486:Y 59342132      21         0 | FALSE
5:2208:14704:146696:Y 59342132      20         0 | FALSE
-----
seqinfo: 25 sequences from hg19 genome

```

3 Session Info

```

> sessionInfo()

R version 3.1.1 Patched (2014-09-24 r66678)
Platform: i386-w64-mingw32/i386 (32-bit)

locale:
[1] LC_COLLATE=C
[2] LC_CTYPE=English_United States.1252
[3] LC_MONETARY=English_United States.1252
[4] LC_NUMERIC=C
[5] LC_TIME=English_United States.1252

attached base packages:

```

```
[1] stats4      parallel  stats      graphics  grDevices  utils      datasets
[8] methods     base
```

other attached packages:

```
[1] RIPSeeker_1.5.0      rtracklayer_1.26.1    GenomicAlignments_1.2.0
[4] Rsamtools_1.18.0     Biostrings_2.34.0     XVector_0.6.0
[7] GenomicRanges_1.18.1 GenomeInfoDb_1.2.0    IRanges_2.0.0
[10] S4Vectors_0.4.0      BiocGenerics_0.12.0
```

loaded via a namespace (and not attached):

```
[1] BBmisc_1.7           BatchJobs_1.4         BiocParallel_1.0.0 DBI_0.3.1
[5] RCurl_1.95-4.3       RSQLite_0.11.4        XML_3.98-1.1      base64enc_0.1-2
[9] bitops_1.0-6         brew_1.0-6            checkmate_1.4      codetools_0.2-9
[13] digest_0.6.4         fail_1.2              foreach_1.4.2      iterators_1.0.7
[17] sendmailR_1.2-1      stringr_0.6.2         tools_3.1.1        zlibbioc_1.12.0
```

References

Jing Zhao, Toshiro K Ohsumi, Johnny T Kung, Yuya Ogawa, Daniel J Grau, Kavitha Sarma, Ji Joon Song, Robert E Kingston, Mark Borowsky, and Jeannie T Lee. Genome-wide Identification of Polycomb-Associated RNAs by RIP-seq. *Molecular Cell*, 40(6):939–953, December 2010.